

Project: 316.0002

Ontwikkeling van (snelle) analysemethoden in het kader van het onderzoekprogramma

"Plantaardige technische oliën als nieuwe landbouwgrondstoffen voor de industrie" - Nationaal  
Olieprogramma

Projectleider: H. Herstel

Rapport 93.21

december 1993

## NIET-DESTRUCTIEVE BEPALING VAN HET OLIEGEHALTE VAN OLIEHOUDENDE ZADEN MET NIR-SPECTROMETRIE

F.P.F. Groenendijk en R. Frankhuizen

Afdeling: Instrumentele Analyse

DLO-Rijks-Kwaliteitsinstituut voor land- en tuinbouwprodukten (RIKILT-DLO)

Bornsesteeg 45, 6708 PD Wageningen

Postbus 230, 6700 AE Wageningen

Telefoon 08370-75400

Telefax 08370-17717

Copyright 1993, DLO-Rijks-Kwaliteitsinstituut voor land- en tuinbouwprodukten.  
Overname van de inhoud is toegestaan, mits met duidelijke bronvermelding.

## VERZENDLIJST

### INTERN:

directeur

auteurs (2x)

projectleider

programmaleiders

afdeling Kwaliteitsbewaking en Kwaliteitssystemen (2x)

afdeling Instrumentele Analyse (4x)

Public relations en Secretariaat (2x)

bibliotheek (3x)

### EXTERN:

Dienst Landbouwkundig Onderzoek, dr. H. Breteler

Directie Milieu, Kwaliteit en Voeding, mr. P. Ritsema

Directie Wetenschap en Technologie, dr. ir. J.M.P. Papenhuijzen

Directie Industrie en Handel, ir. P.J. Bouwmeister

Directie Akker- en Tuinbouw, ir. C.M.M. van Winden

Proefstation voor de Akkerbouw en de Groenteteelt in de Vollegrond, ir. G. Borm

Centrum voor Plantenveredelings- en Reproductie Onderzoek,

ir. L.J.M. van Soest, ir. H. Toxopeus en ir. F. Mulder

Instituut voor Agrobiologisch en Bodemvruchtbaarheidsonderzoek, ir. W. Meijer

Instituut voor Milieu- en Agritechniek, ir. A. Bouman

Leden van NOP-Platform (9x)

## ABSTRACT

Bepaling van het oliegehalte van het zaad van *Crambe abyssinica*, *Euphorbia lagascae*, *Dimorphotheca pluvialis*, *Coriandrum sativum* en *Limnanthes alba* met Nabij Infrarood Reflectie Spectrometrie

Determination of the oil content of the seeds of *Crambe abyssinica*, *Euphorbia lagascae*, *Dimorphotheca pluvialis*, *Coriandrum sativum* and *Limnanthes alba* by Near Infrared Reflection Spectrometry (in Dutch)

Report 93.21

December 1993

F.P.F. Groenendijk en R. Frankhuizen

DLO-State Institute for Quality Control of Agricultural Products (RIKILT-DLO)

PO Box 230, 6700 AE Wageningen, The Netherlands.

10 figures, 13 tables, 8 references, 2 annexes

Near Infrared Reflection Spectrometry (NIRS) was used for rapid nondestructive determination of the oil content of whole seeds of *Crambe abyssinica*, *Euphorbia lagascae*, *Dimorphotheca pluvialis*, *Coriandrum sativum* and *Limnanthes alba*. Because of the inhomogeneity of the seed samples subsamples were made using a Retsch Laboratory Turning Tube Divider. Subsamples were measured using a NIRSystem-6500 instrument equipped with a coarse cuvet or a rotating cuvet. Validation procedures show that calibrations based on seeds of one season are not accurate enough to predict samples of another season. To obtain more robust calibrations, samples should be used of different seasons and growth locations.

Partial least squares (PLS) regression is preferred to multiple linear regression (MLR) because of the speed of the calculations and to avoid colinearity problems.

The oil content of *Crambe*, *Dimorphotheca* and *Coriandrum* were predicted on a two season PLS-calibration with standard errors of prediction (RMSEP) of 0,6 % and *Euphorbia* and *Limnanthes* of 0,4 %. These accuracies are good enough for screening purposes.

Keywords: *Crambe abyssinica*, *Euphorbia lagascae*, *Dimorphotheca pluvialis*, *Coriandrum Sativum*, *Limnanthes alba*, seeds, oil content, Near Infrared Spectrometry, NIR, PLS, MLR



INHOUD	<u>blz</u>
SAMENVATTING	1
1 INLEIDING	3
2 MONSTERMATERIAAL	3
3 METHODEN VAN ONDERZOEK	4
3.1 Monstervoorbereiding	4
3.2 Bepaling van het oliegehalte (klassiek)	4
3.3 NIRS	5
3.3.1 MLR	6
3.3.2 PLS	6
4 RESULTATEN EN DISCUSSIE	7
4.1 Crambe abyssinica	7
4.1.1 Toetsing van de modellen gebaseerd op seizoen 1990	7
4.1.2 Tweede seizoen (oogst 1990 en oogst 1991 samengevoegd)	9
4.2 Euphorbia lagascae	11
4.2.1 Toetsing van de modellen gebaseerd op seizoen 1990	11
4.1.2 Tweede seizoen (oogst 1991)	12
4.3 Dimorphotheca pluvialis	16
4.3.1 Toetsing van de modellen gebaseerd op seizoen 1990	16
4.3.1.1 Kegels	16
4.3.1.2 Vleugels	17
4.3.2 Tweede seizoen (oogst 1991)	17
4.3.2.1 Kegels	17
4.3.2.2 Vleugels	19
4.4 Coriandrum sativum	21
4.5 Limnanthes alba	23
5 CONCLUSIES	24
LITERATUUR	26
BIJLAGEN	
A GEBRUIKTE STATISTISCHE FORMULES	
B TABEL MEEST ROBUUSTE IJKLIJNEN	





## SAMENVATTING

In het kader van het agrificatieprogramma van het Ministerie van LNV "Plantaardige technische oliën als nieuwe landbouwgrondstoffen voor de industrie" zijn de mogelijkheden onderzocht om het oliegehalte van *Crambe abyssinica*, *Euphorbia lagascae*, *Dimorphotheca pluvialis*, *Coriandrum sativum* en *Limnanthes alba* niet-destructief met Nabij Infrarood Reflectie Spectrometrie (NIRS) te bepalen. Het monstermateriaal is afkomstig van veldproeven uitgevoerd in 1990 en 1991. Het oliegehalte is gravimetrisch bepaald door de zaden in petroleum-ether in een kogelschudmolen (Prolabo) te malen. Vervolgens is de olie met petroleum-ether in een Soxhlet-apparaat geëxtraheerd. Het NIRS-onderzoek is uitgevoerd met behulp van een NIRSystem-6500. Van ieder monster is in enkelvoud een NIR-spectrum opgenomen over het golflengtegebied van 1100-2500 nm, waarbij om de 2 nm de reflectie is gemeten in de  $\log 1/R$  ( $R$ =reflectie) vorm. Vervolgens zijn met behulp van meervoudige lineaire regressie-analyse (MLR) en partial least squares regressie (PLS) correlatieberekeningen uitgevoerd tussen de  $\log (1/R)$  waarden behorende bij de verschillende golflengten en de met de referentiemethode bepaalde oliegehalten. Hierbij zijn als belangrijkste criteria de RMSEP (Root Mean Square Error of Prediction) en de  $R$  (multiple correlatiecoëfficiënt) gehanteerd. Berekeningen zijn uitgevoerd met zowel de ruwe spectra als met de 1e en 2e afgeleide spectra. Het is gebleken dat de binnenmonstervariatie van oliehoudende zaden groot is. Gebruik van een grote (coarse) cuvet kan deze variatie uitmiddelen. Deelmonsters nemen met de Retsch-monsterverdeelmolen minimaliseert de variatie verder. De ijklijnen gebaseerd op één enkel seizoen (en/of grondsoort) zijn niet betrouwbaar voor de voorspelling van monsters uit een nieuw seizoen, doordat er verschillen in structuur en samenstelling zijn, die o.a. veroorzaakt worden door seizoensinvloeden, teeltverschillen, grondsoort en afrijpingsstadia. Corrigeren voor bias en/of slope geeft nauwkeurigheden die overeenkomen met die van de oorspronkelijke modellen, maar de verkregen modellen zijn dan alleen aan de nieuwe dataset aangepast. Om meer robuuste ijklijnen te verkrijgen zijn de datasets van beide seizoenen samengevoegd tot één calibratieset. De verwachte nauwkeurigheid (95% betrouwbaarheidsinterval) van de voorspelling van het oliegehalte van monsters uit een volgend seizoen is voor *Crambe abyssinica*, *Dimorphotheca pluvialis* en *Coriandrum sativum* ca. 1,2%, voor *Euphorbia lagascae* ca. 0,9% en voor *Limnanthes alba* ca. 0,8%. Deze nauwkeurigheden zijn voor screeningsdoeleinden voldoende nauwkeurig. Berekeningen met PLS kunnen bij meer complexe matrices een beter resultaat opleveren dan met MLR, maar gemiddeld genomen is het resultaat gelijk. Vanwege de collineariteitsproblemen die bij MLR kunnen optreden en vanwege de snelheid van calibreren gaat de voorkeur uit naar met PLS geconstrueerde ijklijnen.





## 1 INLEIDING

In het kader van het agrificatie-onderzoekprogramma van het Ministerie van LNV "Plantaardige technische oliën als nieuwe landbouwgrondstoffen voor de industrie - Nationaal Olie Programma" (NOP) wordt onderzoek verricht naar een aantal oliehoudende gewassen.

In dit kader is in 1991 een oriënterend onderzoek gestart naar de mogelijkheden om het oliegehalte van *Dimorphotheca pluvialis*, *Crambe abyssinica* en *Euphorbia lagascae* met Nabij Infrarood Reflectie Spectroscopie (NIRS) te bepalen (v.d.Kamp e.a. 1991). Uit de resultaten van dit onderzoek bleek dat NIRS goede mogelijkheden biedt om het oliegehalte van de diverse zaden snel en niet-destructief te bepalen. Wel bestond de indruk dat de nauwkeurigheid van de oliegehaltebepaling in grote mate bepaald wordt door inhomogeniteit van het aangeleverde monstermateriaal.

Onderzoek uitgevoerd in 1992 toont aan dat het splitsen van de monsters in representatieve submonsters met een Retsch-monsterverdeelmolen inderdaad van grote invloed is op de nauwkeurigheid waarmee het oliegehalte zowel klassiek als met NMR bepaald kan worden (v.d.Kamp e.a. 1992). Aangenomen mag worden dat dit eveneens geldt voor de NIR-spectrometrie.

Het hierna beschreven onderzoek had ten doel:

- Toetsen van de betrouwbaarheid van de ijklijnen gebaseerd op monsters uit het seizoen '90 voor *Dimorphotheca*, *Crambe* en *Euphorbia* met monsters uit het seizoen '91.
- Nagaan met welke nauwkeurigheid het oliegehalte in *Coriandrum sativum* en *Limnanthes alba* bepaald kan worden.
- Het kiezen of opstellen van de meest robuuste ijklijnen voor de genoemde zaadtypen.

## 2 MONSTERMATERIAAL

Het monstermateriaal van *Limnanthes* is afkomstig van veldproeven van zowel het Centrum voor Plantenveredelings- en Reproductieonderzoek (CPRO) als van het Proefstation voor de Akkerbouw en de Groenteteelt in de Vollegrond (PAGV). Al het overige monstermateriaal is afkomstig uit veldproeven uitgevoerd door het CPRO.

De *Crambes* zijn afkomstig van een veldproef in 1990 (n=66) en van een veldproef in 1991 (n=34).

De Euphorbia's zijn afkomstig van een veldproef in 1990 (n=75) en van een veldproef in 1991 (n=68) uitgevoerd op de Broekemahoeve (zavel) en gedeeltelijk op de Born (zandgrond).

De Dimorphotheca's zijn van diverse herkomsten en proeven in 1990 (kegels n=35, vleugels n=57) en van een veldproef in 1991 uitgevoerd op de Born (zandgrond, kegels n=23, vleugels n=23).

De monsters Coriandrum zijn afkomstig van een veldproef in 1991 (n=52) uitgevoerd op de Broekemahoeve.

De eerste monsterset Limnanthes is afkomstig van 4 herkomsten die in de winter of in de zomer zijn uitgezaaid in een veldproef in 1991 (CPRO, n=19) uitgevoerd op de Broekemahoeve (klei). De tweede monsterset (CPRO, n=5) is afkomstig van zaadvermeerderingen uitgevoerd op diverse lokaties en diverse jaren (1986-1989). De derde monsterset Limnanthes (PAGV, n=18) is afkomstig van een veldproef in 1991.

### 3 METHODEN VAN ONDERZOEK

#### 3.1 Monstervoorbereiding

Alle monsters zijn gesplitst, tenzij anders vermeld, met een Retsch-monsterverdeelmolen waardoor het monster in 8 representatieve submonsters wordt verdeeld. Zowel voor NIRS als voor de klassieke olie bepaling is met dit apparaat een submonster gemaakt met het juiste gewicht. Dit betekende dat de opsplitsing 2 tot 3 maal op hetzelfde monster moest worden uitgevoerd tot het juiste gewicht was bereikt (eventueel m.b.v. meerdere submonsters).

#### 3.2 Bepaling van het oliegehalte (klassiek)

Van de monsters is in enkelvoud het oliegehalte gravimetrisch bepaald conform intern voorschrift A 104. De zaden worden in petroleum-ether in een kogelschudmolen (Prolabo) vermalen. Vervolgens wordt met petroleum-ether de olie geëxtraheerd, het oplosmiddel afgedampt en het oliegehalte bepaald na drogen bij 80°C onder vacuüm tot constant gewicht.



### 3.3 NIRS

Het NIRS-onderzoek is uitgevoerd met behulp van een NIRSystem-6500. Het betreft hier een modulair opgebouwde spectrofotometer, waarmee zowel in reflectie als in transmissie gemeten kan worden over een breed golflengtegebied (400-2500 nm). Bij Euphorbia-zaadmonsters kan olie uittreden waardoor de cuvet na elke meting ontvet moet worden. Voor de Euphorbia-zaadmonsters is daarom gekozen voor reflectiemetingen in een roterende monstercup, (benodigde hoeveelheid ca. 8 ml zaad) omdat deze cups makkelijk schoon te maken zijn na gebruik. Bij deze monstercups wordt van het totale oppervlak ca. 4.5 cm<sup>2</sup> aangestraald. De overige zaadmonsters zijn gemeten in een coarse-cuvet die met minimaal 30 ml zaad en maximaal met 100 ml zaad gevuld kan worden. Het grote voordeel van deze cuvet is dat er aan een grote hoeveelheid monster gemeten wordt waardoor inhomogeniteit van het monster uitgemiddeld wordt. Het effectief oppervlak voor de coarse cuvet is ca. 70 cm<sup>2</sup>.

Van ieder monster is in enkelvoud een NIR-spectrum opgenomen (Euphorbia in duplo) over het golflengtegebied van 1100-2500 nm, waarbij om de 2 nm de reflectie is gemeten in de log 1/R (R=reflectie) vorm. Als referentie (100% reflectie) is een keramisch schijfje gebruikt.

Systeemcontrole vond plaats conform de procedure beschreven in de NSAS software (release 3.25).

Van de monsters zijn de met de referentiemethode bepaalde oliegehalten gekoppeld aan de bijbehorende NIR-spectra. Met behulp van meervoudige lineaire regressie-analyse (MLR) en Partial Least Squares (PLS) zijn voorspelmodellen (ijklijnen) opgesteld voor de bepaling van het oliegehalte (zie bijlage A voor de gebruikte statistische formules). De eerder opgestelde ijklijnen gebaseerd op monsters uit het seizoen 1990 zijn op hun betrouwbaarheid getoetst met behulp van een toetsset met monsters uit het seizoen 1991. Daarna zijn de calibratiesets van *Dimorphotheca*, *Crambe* en *Euphorbia* van de beide seizoenen (1990 en 1991) samengevoegd en zijn er voor deze zaadtypen nieuwe ijklijnen opgesteld. Tot slot zijn voor *Coriandrum sativum* en *Limnanthes alba* met behulp van monsters uit het seizoen 1991 voorspelmodellen opgesteld. Indien mogelijk zijn alle opgestelde modellen getoetst met monsters die geen deel uit maken van de calibratieset. Als belangrijkste maat voor de nauwkeurigheid van de ijklijnen is gekozen voor de standaardafwijking van de verschillen tussen de met NIRS voorspelde en de met de klassieke methode gevonden oliegehalten van de toetsmonsters (Root Mean Square Error of Prediction (RMSEP)). Indien de RMSEP van de zelfde orde van grootte is als de RMSEC en indien de intercept en slope van de regressielijn van de toetsset niet significant afwijken van die van de ijklijn, wordt de ijklijn betrouwbaar geacht.

### 3.3.1 MLR

Meervoudige lineaire regressie-analyse (MLR) is een standaardmethode voor het aanpassen van lineaire verbanden (Montgomery & Peck 1982, Draper & Smith 1981). De methode is echter niet toepasbaar als het aantal voorspellende variabelen groot is (niet groter dan 1/3 van het aantal eenheden in de dataset). Bovendien leiden hoge correlaties tussen de predictoren tot instabiele modellen (collineariteitsprobleem). Om deze redenen is MLR niet geschikt voor spectrale data.

Wel kan men MLR toepassen in combinatie met een methode voor golflengteselectie. De voorspellingen worden dan niet gebaseerd op het gehele spectrum, maar slechts op een beperkt aantal golflengten. Het zoeken naar de meest significante golflengten en de beste combinaties is een tijdrovende bezigheid die veelal gebaseerd is op een combinatie van statistische grootheden en de trial and error methode. In dit onderzoek zijn met behulp van MLR correlaties berekend tussen de "klassiek" bepaalde oliegehalten en de bij de verschillende golflengten verkregen NIR-reflectiewaarden. Hierbij zijn golflengten geselecteerd en calibratiefactoren berekend die in combinatie de hoogste meervoudige correlatiecoëfficiënt ( $R$ ) en de kleinste standaardafwijking van de verschillen ( $RMSEC = \text{Root Mean Square Error of Calibration}$ ) gaven tussen de bepaalde oliegehalten en de met NIRS berekende gehalten. Er zijn zowel correlatieberekeningen uitgevoerd aan de ruwe spectra als aan de 1e en 2e afgeleide spectra (bij afgeleide spectra wordt er gecorrigeerd voor eventuele verschillen in reflectie veroorzaakt door verschillen in structuur zoals o.a. deeltjesgrootte). Tevens is er bij de afgeleide spectra gerekend met verschillende segment- en gapgrootten ter verkrijging van een optimale signaal/ruisverhouding.

### 3.3.2 PLS

Partial Least Squares (PLS) regressie is een nog niet zo oude, maar tegenwoordig populaire calibratiemethode voor hoogdimensionale datasets (Martens & Naes 1989). Implementaties van PLS zijn o.a. beschikbaar in het pakket NSAS (NSAS 1991), in het pakket Unscrambler (Unscrambler 1992) en in de algemene statistische programmeertaal Genstat (Genstat 1987).

Met PLS wordt een lineair model geconstrueerd tussen een te voorspellen variabele  $y$  en de voorspellers  $x_j$  ( $j = 1 \dots p$ ). Om de collineariteitsproblemen die bij MLR optreden te vermijden, worden een aantal latente variabelen  $t_k$  ( $k = 1 \dots a$ ) geconstrueerd die men zou kunnen zien als een beschrijving van die variatie in de spectra die relevant is voor het voorspellen van  $y$ . Hierbij is het aantal latente variabelen ( $a$ ) meestal veel kleiner dan het aantal predictoren  $p$ . Bovendien zijn de latente variabelen  $t_k$  onderling niet gecorreleerd. Het voorspellen van  $y$  uit  $t_1, \dots, t_a$  gebeurt dan ook met gewone lineaire regressie.



Kenmerkend voor PLS is de wijze waarop de latente variabelen  $t_k$  tot stand komen. De latente variabelen worden één voor één geconstrueerd volgens een eenvoudig principe:

Elke latente variabele  $t_k$  is evenredig met een gewogen gemiddelde van alle variabelen  $x_1 \dots x_p$ , waarbij de weegfactoren de covarianties tussen  $y$  en  $x_j$  ( $j = 1 \dots p$ ) zijn, voor zover deze informatie nog niet in eerdere latente variabelen is benut. Een variabele  $x_j$  die een relatief hoge correlatie heeft met  $y$ , zal dus sterk meewegen in de eerste latente variabele.

In sommige PLS uitvoeringen worden alle variabelen  $x_j$  van te voren nog geschaald tot een gelijke variantie, zodat alle golflengten a priori hetzelfde belang krijgen. In het hier beschreven onderzoek is hiervan geen gebruik gemaakt. Bij het gebruik van PLS dient de dimensionaliteit ( $a$ ) van het model gekozen te worden. Dit is gebeurd door voor alle waarden tussen  $a = 1$  en (meestal)  $a = 15$  de voorspelfout te schatten, en  $a$  zodanig te kiezen dat deze voorspelfout minimaal wordt. De voorspelfout is berekend als de Root Mean Squared Error of Prediction (RMSEP) via de cross-validatie methode.

## 4 RESULTATEN EN DISCUSSIE

### 4.1 Crambe abyssinica

#### 4.1.1 Toetsing van de modellen gebaseerd op seizoen 1990

Aan de hand van niet gesplitste monsters ( $n=66$ , oogst 1990) zijn rekenmodellen opgesteld. Deze zijn getoetst m.b.v. de monsterset ( $n=34$ ) van het seizoen 1991. In tabel 4.1 zijn deze toetsresultaten weergegeven. Hieruit blijkt dat er zowel een grote bias-correctie als een grote slope-correctie nodig is om het tweede seizoen met voldoende nauwkeurigheid te voorspellen. Door te corrigeren voor bias ( $RMSEP_{bc}$ ) en slope ( $RMSEP_{bsc}$ ) kan een schatting worden verkregen van de haalbare RMSEP. Eén ijklijn (C08T3F3A) lijkt te voldoen voor beide seizoenen.

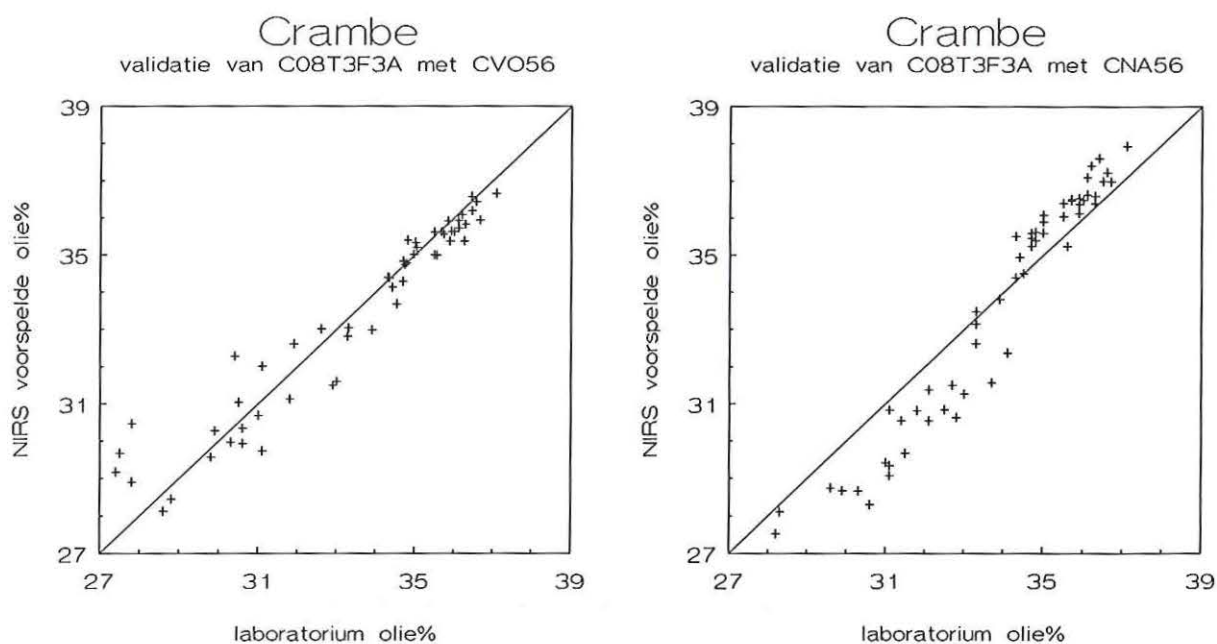
Tabel 4.1 Resultaten van de toetsing van de ijklijnen gebaseerd op monsters uit oogst 1990 met de monsters van oogst 1991 (n=34).

Ijklijn	RMSEC	RMSEP	Bias	Slope	RMSEP <sub>bc</sub>	RMSEP <sub>bsc</sub>
C08-3T0F	1,1	6	6	0,7	0,62	0,43
C08-3T2F	1,1	4	4	0,6	0,73	0,55
C08T3F3A	1,1	0,42	-0,23	0,91	0,35	0,34
C083T3FC	1,1	2	2	0,7	0,57	0,32
C88R3TF2	0,9	5	5	0,6	0,73	0,55

Geconcludeerd kan worden dat modellen gebaseerd op één oogstjaar niet zonder meer het juiste niveau voorspellen van monsters uit een nieuw oogstjaar, hetgeen zich uit in een bias en/of een verschil in slope. Tijdens onderzoek in 1991 (v.d.Kamp e.a. 1991) bleek dat de invloed van de homogeniteit van de monsters op de nauwkeurigheid van de voorspelling groot is. Omdat de oorspronkelijke cijfers afkomstig zijn van niet opgesplitst materiaal (inhomogeen), is besloten om de monsters opnieuw in te lezen en het oliegehalte opnieuw te bepalen. In december 1991 zijn 22 monsters uit 1990 en 34 monsters uit 1991 opnieuw ingelezen. Dit resulteert in een nieuwe dataset CNA9091 (n=56). Van de monsters uit 1990 is het oliegehalte opnieuw klassiek bepaald. Zowel voor de klassieke bepaling als voor de inlezing zijn de monsters gesplitst op de Retsch-monsterverdeel-molen.

In figuur 4.1 zijn de oliegehalten van deze Crambe-monsters bepaald met de NIRS-ijklijn C08T3F3A uitgezet tegen de klassiek bepaalde gehalten. Links (CVO56) is de oorspronkelijke inlezing (december 1990 en oktober 1991) weergegeven en rechts (CNA56) dezelfde opnieuw ingelezen monsters (december 1991). De Retsch-gesplitste monsters uit 1990 (figuur 4.1, rechts; monsters 27%-34%) hebben een kleinere RMSEP (0,6% versus 1,2%) terwijl er ook gemiddeld 1,4 % meer olie in gevonden wordt dan in de niet gesplitste monsters. De nieuw ingelezen monsters uit 1991 (figuur 4.1, rechts monsters 34%-39%) worden t.o.v. de ijklijn te hoog (0,6%) voorspeld. Uit deze verschillen blijkt dat de ijklijn niet robuust is. Er is terecht besloten om de monsters opnieuw in te lezen. Met de opnieuw ingelezen monsters zijn nieuwe modellen opgesteld (§4.1.2).

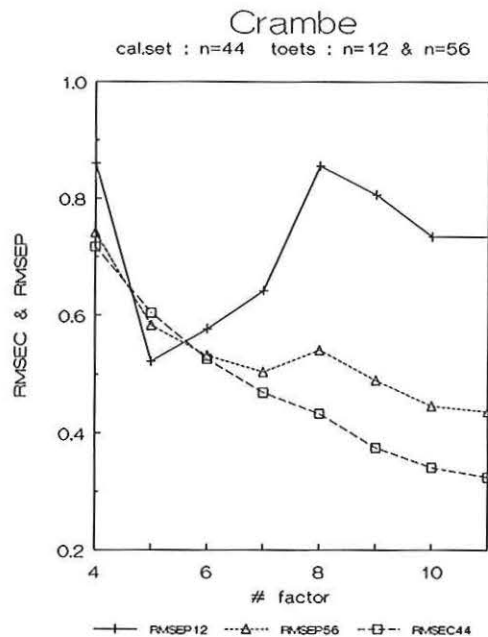




Figuur 4.1 Voorspelling op de ijklijn C08T3F3A van de samengevoegde monsterset (1990 + 1991) gesplitst in een oorspronkelijke inlezing (CVO56,links) en een nieuwe inlezing (CNA56,rechts).

#### 4.1.2 Tweede seizoen (oogst 1990 en oogst 1991 samengevoegd)

De nieuwe dataset (CNA9091,  $n=56$ ) is verdeeld in een calibratieset ( $n=44$ ) en een toetsset ( $n=12$ ), door selectie over de range. Voor het opzetten van de rekenmodellen is gebruik gemaakt van MLR en PLS. Met de calibratieset zijn verschillende modellen opgesteld, die vervolgens zijn getoetst met de toetsset. In figuur 4.2 is het aantal PLS-factoren uitgezet tegen de RMSEC, de RMSEP van de hele set ( $n=56$ ) en de toetsset ( $n=12$ ). Bij de zesde factor neemt de RMSEP weer toe, wat duidt op overfit. Hieruit volgt dat het optimale aantal factoren vijf (5) is.



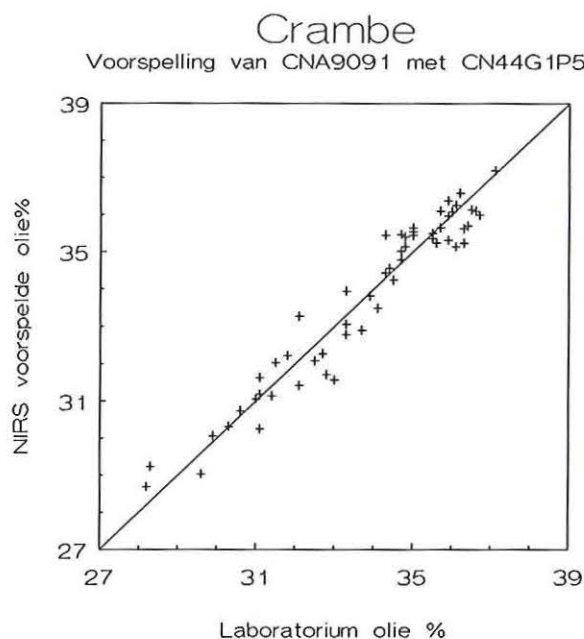
Figuur 4.2 Het aantal PLS-factoren uitgezet tegen de RMSEC (n=44) en de RMSEP van de hele set (n=56) en van de toetsset (n=12).

In tabel 4.2 zijn de ijklijnen opgenomen, die het meeste perspectief bieden. Voor de ijklijn C08T3F3A (oogst 1990) geldt dat de in de tabel vermelde RMSEP pas bereikt wordt na correctie voor slope en bias aan de hand van de nieuwe dataset (n=56) en de toetsset (n=12).

Tabel 4.2 Gegevens van de Crambe ijklijnen die het meeste perspectief bieden. Voor de laatste ijklijn (oogst 1990) is de RMSEP voor bias en slope gecorrigeerd.

ijklijn	methode	n	RMSEC	RMSEP <sup>56</sup>	RMSEP <sup>12</sup>	Bias <sup>12</sup>	Slope <sup>12</sup>
C56NT32F	MLR	56	0,49	0,49	0,57	-0,07	0,895
C4N2T3F3	MLR	40	0,50	0,57	0,54	-0,11	0,970
CN56P2S5	PLS	56	0,43	0,43	0,46	0,00	0,921
CN44G1P5	PLS	44	0,60	0,58	0,52	-0,20	0,928
C56P5	PLS	56	0,60	0,60	0,41	-0,14	0,983
C08T3F3A	M <sup>'90</sup>	66 <sup>'90</sup>	1,10	0,55c	0,53c	-0,67	0,659

De ijklijnen geven een RMSEP van ca. 0,6%. Dit wil zeggen dat het oliegehalte van Crambe bepaald kan worden met NIRS met een nauwkeurigheid van ca. 1,2% (95% betrouwbaarheid). In figuur 4.3 zijn van een ijklijn (CN44G1P5) de oliegehalten die klassiek zijn bepaald uitgezet tegen de met NIRS voorspelde oliegehalten (voorspeld aan zichzelf).



Figuur 4.3 Oliegehalten van Crambe ( $n=44$ ) bepaald met NIRS m.b.v. de ijklijn CN44G1P5 uitgezet tegen de klassiek bepaalde oliegehalten.

## 4.2 *Euphorbia lagascae*

### 4.2.1 Toetsing van de modellen gebaseerd op seizoen 1990

De rekenmodellen uit het seizoen 1990 zijn opgesteld aan de hand van niet gesplitste monsters ( $n=75$ ). De berekende RMSEC bedraagt 0,65%, waardoor de verwachte nauwkeurigheid ca. 1,3% (95% betrouwbaarheidsinterval) bedraagt. Deze rekenmodellen zijn getoetst m.b.v. de monsterset ( $n=68$ ) van het seizoen 1991. De monsters uit 1991 zijn zowel voor de klassieke bepaling als voor de NIRS-meting gesplitst m.b.v. de Retsch-monsterverdeelmolen. Op de dataset uit 1990 is ook gerekend met PLS. In tabel 4.3 zijn de toetsresultaten opgenomen. De gevonden RMSEP's zijn relatief hoog t.o.v. de RMSEC's het geen veroorzaakt wordt door een bias tussen de gehalten van de monsters uit beide seizoenen. Door te corrigeren voor de bias ( $RMSEP_{bc}$ ) en de slope ( $RMSEP_{bsc}$ ) kan een schatting worden verkregen van de haalbare RMSEP. Het is mogelijk om een RMSEP van ca. 0,7% met MLR en ca. 0,5% met PLS te bereiken met een nauwkeurigheid van respectievelijk 1,4% en 1,0% (95% betrouwbaarheidsinterval). Met de correcties wordt de ijklijn aangepast aan het nieuwe seizoen. Het is echter aan te bevelen beide seizoenen samen te voegen ( $n=143$ ) en een nieuwe ijklijn te berekenen, waardoor alle eigenschappen in de calibratie meegenomen worden en de ijklijn robuuster wordt.

Tabel 4.3 Resultaten van de toetsing van de ijklijnen uit 1990 met de data uit 1991. Door correctie voor bias wordt de  $RMSEP_{bc}$  verkregen en door correctie voor bias en slope wordt  $RMSEP_{bsc}$  verkregen.

ijklijn	methode	RMSEC	$RMSEP'^{91}$	bias	slope	$RMSEP_{bc}$	$RMSEP_{bsc}$
E7-2T0FA	MLR	0,64	0,82	0,44	1,136	0,70	0,68
E7-2T1FA	MLR	0,65	0,83	0,24	0,942	0,80	0,79
E7-3T2FA	MLR	0,64	1,00	0,69	1,028	0,73	0,73
EM08	PLS	0,59	0,79	0,60	0,826	0,51	0,44
EM05	PLS	0,65	0,92	0,81	0,897	0,42	0,39
E75G15P5	PLS	0,64	0,75	0,56	1,037	0,48	0,48
E75G15P6	PLS	0,58	1,03	0,90	1,015	0,49	0,49

#### 4.2.2 Tweede seizoen (oogst 1991).

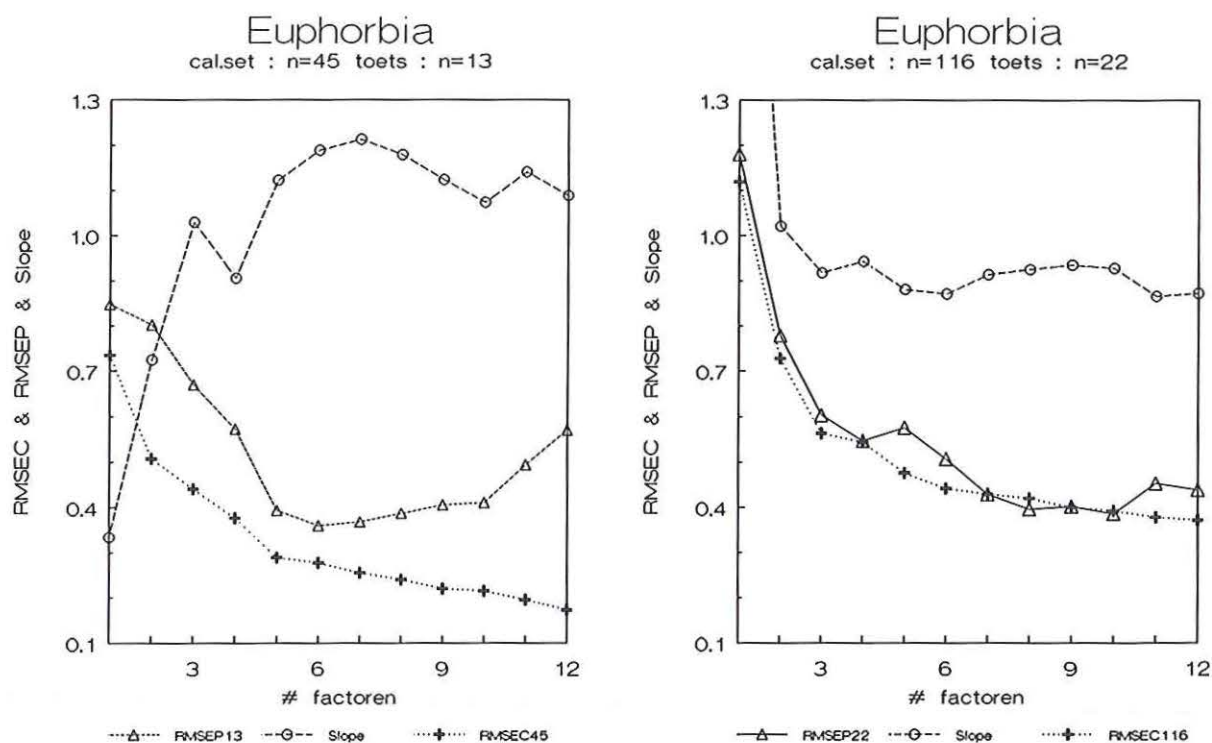
De monsterset van het seizoen 1991 ( $n=68$ ) bevat een subset ( $n=10$ ) die een lager oliegehalte heeft dan de overige monsters ( $n=58$ ). Deze verschillen worden veroorzaakt doordat de monsters uit deze subset op zand zijn geteeld en de overigen op klei. In een deel van de berekeningen is deze subset meegenomen in de calibratie, omdat ook de range van het oliegehalte van de set uit 1990 grotendeels lager ligt dan die van de set uit 1991. De dataset is opgesplitst in een calibratieset ( $n=50$  inclusief zandteelt monsters,  $n=45$  exclusief zandteelt) en een toetsset ( $n=18$  inclusief,  $n=13$  exclusief). In figuur 4.4 (links) is voor de calibratieset ( $n=45$ ) het aantal PLS-factoren uitgezet tegen de RMSEP van de toetsset ( $n=13$ ). Gekozen is voor het model met 5 factoren, omdat bij meer factoren de slope toeneemt en daardoor verder afwijkt van het ideaal (1.000). In tabel 4.4 zijn de ijklijnen opgenomen die het meeste perspectief bieden. Het is mogelijk om een RMSEP van ca. 0,4% te bereiken met een nauwkeurigheid van 0,8% (95% betrouwbaarheidsinterval). De zandteeltmonsters beïnvloeden de ijklijn onevenredig (slope=1,3 en bias=0,4) en moeten niet in de calibratie worden meegenomen als het model gebruikt wordt om monsters uit 1991 te voorspellen.



Tabel 4.4 Gegevens van de Euphorbia-ijklijnen van het seizoen 1991 die het meeste perspectief bieden. De vermelde toetsset  $n=10_z$  bestaat volledig uit de zandteeltmonsters.

ijklijn	methode	RMSEC	n-toets	RMSEP	bias	slope
E5AU4280	MLR	0,34	18	0,43	-0,07	1,005
E0650G5	PLS	0,32	18	0,33	-0,06	0,973
E45-5	PLS	0,29	13	0,39	-0,05	1,123
E45-5	PLS	0,29	$10_z$	0,67	0,41	1,299

De datasets uit 1990 en 1991 zijn vervolgens samengevoegd tot één nieuwe set, waardoor de seizoensinvloeden in de nieuwe ijklijnen worden meegenomen. De samengestelde dataset (1990 + 1991,  $n=143$ ) is verdeeld in een calibratieset ( $n=119$ ) en een toetsset ( $n=24$ ), door selectie over de range. Daarnaast is er van uitgegaan dat voor de niet-Retsch-gesplitste set (1990) geldt dat de grootste afwijking wordt veroorzaakt door niet-representatieve monsternamen. De (5) monsters die een duidelijk grotere afwijking in de voorspelling vertonen dan de rest van de set zijn uit de set verwijderd. Ook met deze dataset ( $n=138$ ) zijn modellen opgesteld aan de hand van een calibratieset ( $n=116$ ) en een toetsset ( $n=22$ ). In figuur 4.4 (rechts) zijn de PLS-factoren uitgezet tegen de RMSEP van de toetsset ( $n=22$ ). In het RMSEP-optimum (8 factoren) is de RMSEP slechts een fractie (0,03%) kleiner dan bij 7 factoren. Gelet op het kleine aantal monsters en het relatief grote aantal factoren is gekozen voor 7 factoren om overfit te voorkomen.



Figuur 4.4 Het aantal PLS-factoren uitgezet tegen de RMSEP, slope en RMSEC van het tweede seizoen (n=58, links) en van de twee seizoenen samen (n=138, rechts).

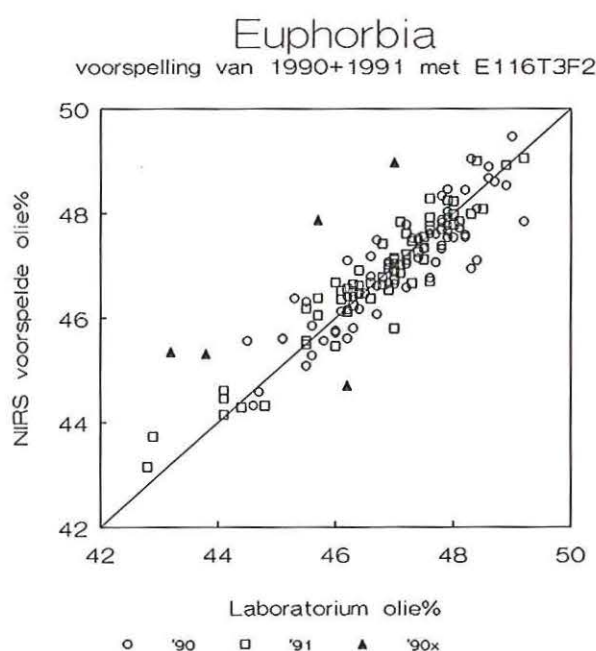
In tabel 4.5 zijn de ijklijnen opgenomen die het meeste perspectief bieden. Deze ijklijnen hebben in de voorspelling een verwachte nauwkeurigheid van ca. 0,8-1,0% (95% betrouwbaarheids- interval). Om de invloed van de zandteelt-subset op de ijklijn gebaseerd op beide seizoenen te bepalen, is ook gerekend zonder deze monsters (calibratieset n=109, toetsset n=19). De resulterende ijklijn (RMSEC=0,43, RMSEP=0,47) voorspelt de zandteeltsubset (n=10) op gelijkwaardig niveau (RMSEP=0,51, Bias=0,16 en Slope=1,017). Hieruit is geconcludeerd dat de invloed van de subset klein is en niet bepalend voor de ligging van de ijklijn. De extra informatie van deze monsters zal mogelijk de ijklijn robuuster maken.



Tabel 4.5 Gegevens van de Euphorbia ijklijnen die het meeste perspectief bieden.

ijklijn	seizoen	n-cal	methode	RMSEC	RMSEP <sup>22</sup>	bias	slope
E5AU4280	'91	58	MLR	0,34	0,47	0,07	0,974
ED19MLR3	'90+'91	119	MLR	0,57	0,40	-0,08	0,974
E116T3F2	'90+'91	116	MLR	0,47	0,40	0,01	0,914
E45-5	'91	45	PLS	0,29	0,43	-0,03	1,089
E0650G15	'91	50	PLS	0,32	0,43	-0,20	0,977
ET06	'90+'91	143	PLS	0,57	0,48	-0,03	0,919
ETR06	'90+'91	138	PLS	0,44	0,45	0,03	0,910
E11607	'90+'91	116	PLS	0,43	0,43	0,04	0,914

In figuur 4.5 is een voorspelling met behulp van de ijklijn E116T3F2 weergegeven van de beide seizoenen (1990 en 1991). Hierbij zijn de oliegehalten die klassiek zijn bepaald uitgezet tegen de met NIRS voorspelde oliegehalten (voorspeld aan zichzelf). De monsters ('90x) met een afwijkend oliegehalte, dat veroorzaakt is door slechte deelmonsters uit een inhomogeen bulkmonster, zijn apart opgenomen.



Figuur 4.5 Voorspelling van de seizoenen 1990 en 1991 op de ijklijn E116T3F2. De monsters met een afwijkend oliegehalte, dat veroorzaakt is door slechte deelmonsters uit een inhomogeen bulkmonster, zijn apart opgenomen ('90x).

### 4.3 Dimorphotheca pluvialis

#### 4.3.1 Toetsing van de modellen gebaseerd op seizoen 1990

De monsters uit het seizoen 1990 zijn niet opgesplitst. De modellen uit 1990 zijn met de datasets uit 1991 getoetst. Uit de toets blijkt één subdataset (zowel voor kegels als voor vleugels  $n=15$ ) het bijzonder slecht te doen. Bij navraag bleek dat deze monsterset lichtelijk beschimmeld is geraakt tijdens de opslag. Waarschijnlijk is de oppervlaktestructuur veranderd waardoor de spectra niet overeenkomen met de overige monsters. Een model opgesteld voor deze ( $n=15$ ) monsters alleen, vertoont wel een goede correlatie. De modellen voor Dimorphotheca zijn zowel met MLR als met PLS opgesteld. De beschimmelde monsters zijn niet opgenomen in de calibratie.

##### 4.3.1.1 Kegels

In tabel 4.6 zijn de modellen opgenomen die het meeste perspectief bieden. Met MLR en PLS wordt op grond van oogst '90 een voorspellingsnauwkeurigheid verwacht van ca. 1,8% respectievelijk ca. 1,2% (95% betrouwbaarheidsinterval). De nauwkeurigheid van de voorspelling blijkt echter slechter dan aan de hand van de RMSEC verwacht mocht worden, maar dit kan worden toegeschreven aan het verschil tussen de twee seizoenen (afgaande op de bias). Het is beter om het nieuwe seizoen aan de calibratie toe te voegen, zodat alle eigenschappen in de nieuwe lijn worden meegenomen en deze robuuster wordt.

Tabel 4.6 Resultaten van de toetsing van de ijklijnen uit 1990 met de data uit 1991. Door correctie voor bias wordt de  $RMSEP_{bc}$  verkregen en door correctie voor bias en slope wordt  $RMSEP_{bsc}$  verkregen.

ijklijn	methode	RMSEC <sup>'90</sup>	RMSEP <sup>'91</sup>	bias	slope	$RMSEP_{bc}$	$RMSEP_{bsc}$
DK80T3F2	MLR	0,89	1,20	-0,34	1,310	1,10	0,90
DKMD23-7	PLS	0,71	0,96	-0,45	0,917	0,84	0,80
DKMD-2-7	PLS	0,57	0,97	0,58	0,969	0,77	0,76
DKMD-4-6	PLS	0,59	1,20	0,99	0,971	0,73	0,72
DKMA6	PLS	0,91	0,94	0,56	1,120	0,74	0,67

#### 4.3.1.2 Vleugels

Er zijn een aantal modellen opgesteld. In tabel 4.7 zijn de modellen opgenomen die het meeste perspectief bieden. Met MLR en PLS wordt een voorspellingsnauwkeurigheid verwacht van respectievelijk ca. 2,0% en ca. 1,6% (95% betrouwbaarheidsinterval). De nauwkeurigheid blijkt echter iets slechter dan aan de hand van de RMSEC verwacht mocht worden, maar net als bij de kegels kan dit worden toegeschreven aan het verschil tussen de twee seizoenen (afgaande op de bias). Ook hier is het beter om het nieuwe seizoen aan de calibratie toe te voegen, zodat alle eigenschappen in de nieuwe lijn worden meegenomen en deze robuuster wordt.

Tabel 4.7 Resultaten van de toetsing van de ijklijnen uit 1990 (n=57) met de data uit 1991 (n=23). Door correctie voor bias wordt de  $RMSEP_{bc}$  verkregen en door correctie voor bias en slope wordt  $RMSEP_{bsc}$  verkregen.

ijklijn	methode	RMSEC <sup>'90</sup>	RMSEP <sup>'91</sup>	bias	slope	RMSEP <sub>bc</sub>	RMSEP <sub>bsc</sub>
DN4421T3	MLR	1,0	2,2	1,8	1,02	1,2	1,2
DN4422T3	MLR	0,92	0,95	-0,41	1,08	0,85	0,84
D5705	PLS	1,0	1,3	0,36	0,74	1,2	1,1
DN44G1P6	PLS	0,77	0,91	-0,60	1,05	0,68	0,67
D5722-7	PLS	0,76	0,86	-0,47	1,01	0,71	0,71

#### 4.3.2 Het tweede seizoen (oogst 1991)

Van het tweede seizoen (oogst 1991) zijn, zowel voor de vleugels als voor de kegels, slechts 23 monsters beschikbaar. Deze datasets zijn niet opgedeeld in een calibratieset en een toetsset, omdat de sets nu al bijna te klein zijn om een betrouwbaar model op te stellen.

##### 4.3.2.1 Kegels

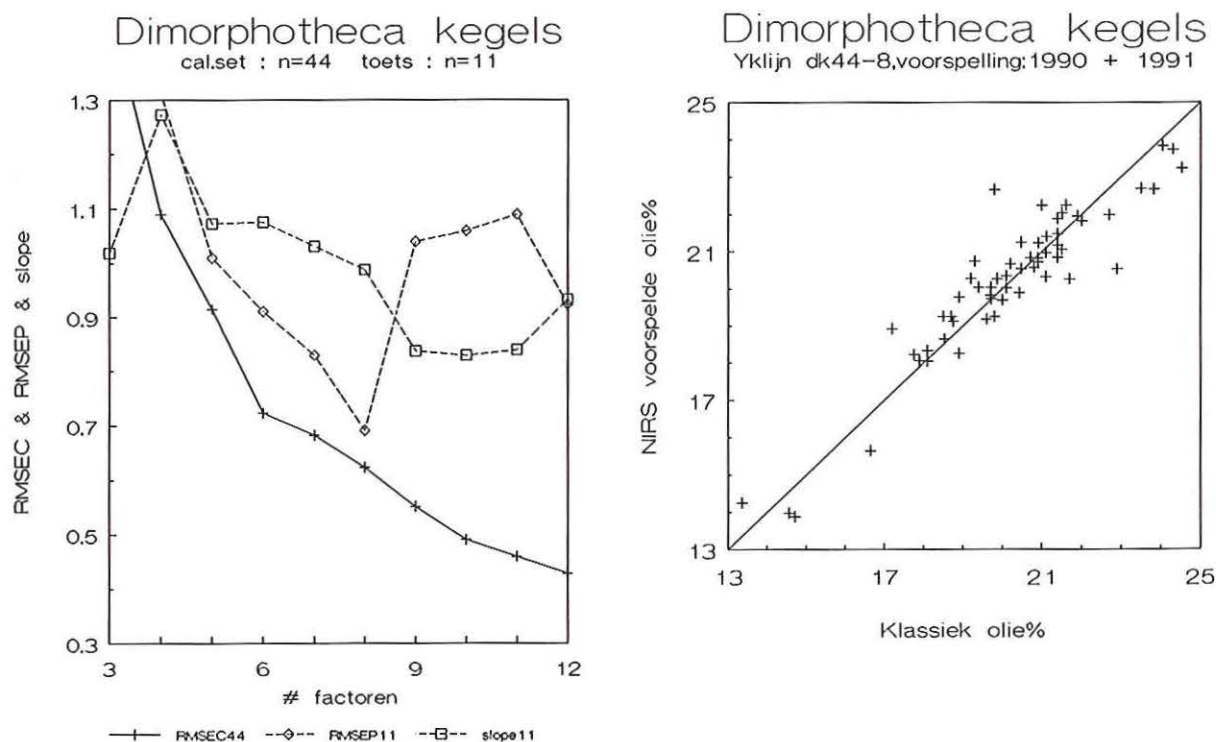
In tabel 4.8 zijn de ijklijnen opgenomen die voor deze beperkte set het meeste perspectief bieden. De verwachte RMSEP (uit de gecorrigeerde RMSEP §4.3.1.1) is ca. 1,4-1,8%. In de toets is te zien dat er met deze ijklijnen nagenoeg geen verband wordt gevonden met het seizoen 1990. Het aantal monsters (n=23) is ook te klein om een acceptabel model te maken.



Tabel 4.8 Gegevens van de Dimorphotheca-kegel-ijklijnen van het seizoen 1991 (n=23) die het meeste perspectief bieden.

ijklijn	methode	RMSEC <sup>'91</sup>	RMSEP <sup>'90</sup>	bias	slope	RMSEP <sub>bc</sub>	RMSEP <sub>bsc</sub>
DK3T3F92	MLR	0,68	2,25	-1,73	0,614	1,4	1,2
DKG5P6	PLS	0,45	2,80	-1,77	0,397	2,1	1,2

Het samenvoegen van beide seizoenen geeft een set (n=58) met de eigenschappen van beide seizoenen. Er is gerekend met de hele set en met een opgedeelde set (calibratie n=44, toets n=11), waarbij de monsters met constante grote afwijking (n=3) uit de berekeningen zijn gelaten, omdat dit wordt toegeschreven (bij een afwijking van meer dan drie maal de RMSEC) aan de monsternamen. In figuur 4.6 (links) is het aantal PLS-factoren van de calibratieset (n=44) uitgezet tegen de RMSEC, RMSEP en slope uit de voorspelling van de toetsset (n=11). Uit het optimum (minimum) van de RMSEP is het juiste aantal factoren (8) bepaald. Ter bevestiging van deze keuze is gebruik gemaakt van de curve van de slope. De bias is niet opgenomen, omdat deze in dit geval geen aanvullende informatie verschaft. Er is gekozen voor acht (8) factoren. Tevens is in figuur 4.6 (rechts) de voorspelling opgenomen van de hele set (n=58) met de ijklijn DK44-8 (met 8 PLS-factoren).



Figuur 4.6 Het aantal PLS-factoren uitgezet tegen de RMSEC, RMSEP en slope uit de voorspelling van de toetsset (n=11, links). Oliegehalten van Dimorphotheca-kegels bepaalt met NIRS m.b.v. de ijklijn DK44-8 (8 PLS-factoren), voorspelling aan de hele set (n=58, rechts).

In tabel 4.9 zijn de ijklijnen opgenomen die het meeste perspectief bieden. Alleen de calibratieset met  $n=44$  is onafhankelijk getoetst ( $n=11$ ). Bij de overige modellen is de toetsset opgenomen in de calibratieset, waardoor de RMSEP-waarden indicatief zijn. De voorspelling heeft een nauwkeurigheid van ca. 1,2% met PLS en ca. 1,5% met MLR (95% betrouwbaarheidsinterval). Deze nauwkeurigheid is iets beter dan uit de gecorrigeerde RMSEP (§4.3.1.1) verwacht mocht worden.

Tabel 4.9 Gegevens van de Dimorphotheca-kegel-ijklijnen die het meeste perspectief bieden. Alleen de calibratieset met  $n=44$  is onafhankelijk getoetst ( $n=11$ ).

ijklijn	methode	RMSEC	R	n-cal	RMSEP <sup>11</sup>
DK90914T	MLR	0,75	0,94	58	0,67
DKT14F1	MLR	0,72	0,95	58	0,74
DK44-8	PLS	0,62	0,96	44	0,69
DKT7-7	PLS	0,63	0,94	58	0,55
DKT2A-6	PLS	0,76	0,94	58	0,75

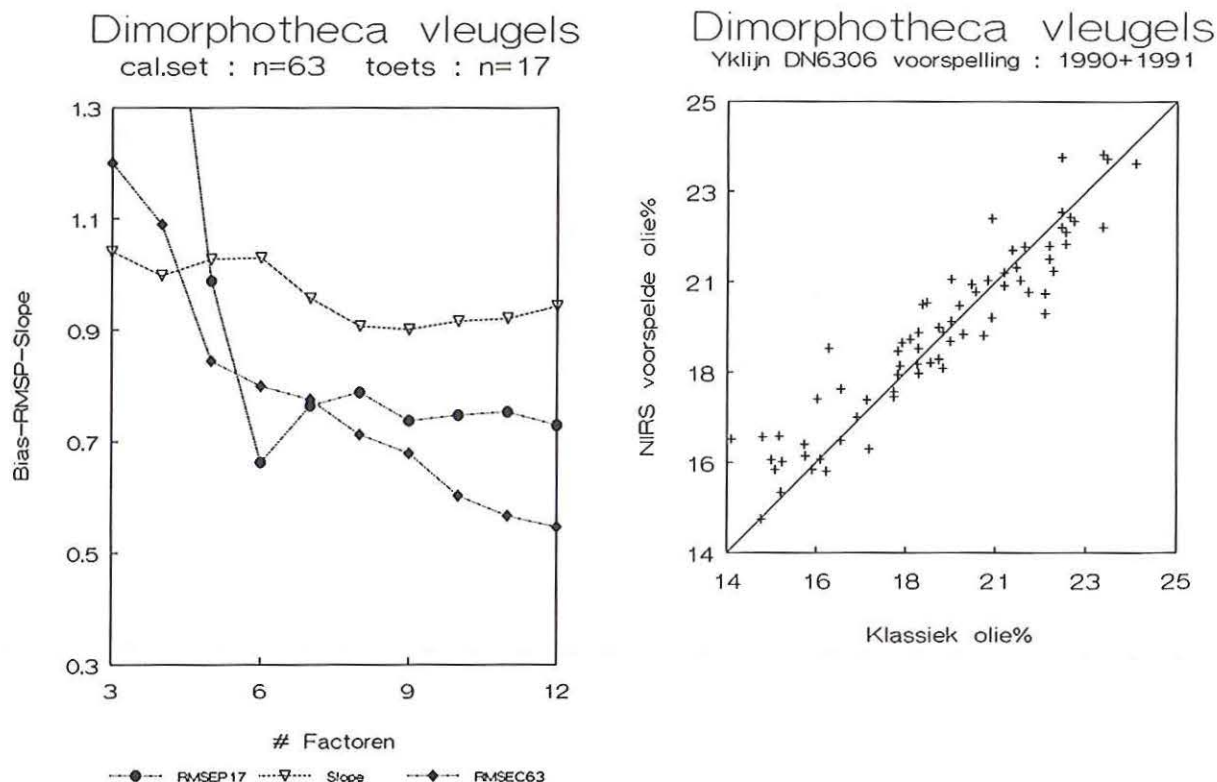
#### 4.3.2.2 Vleugels

Ook voor de vleugels geldt dat de set ( $n=23$ ) te klein is voor een redelijk calibratiemodel. De modellen kunnen slechts een indicatie geven van het verwachtingspatroon. In tabel 4.10 staan enkele perspectiefrijke voorbeelden. De verwachte RMSEP (uit de gecorrigeerde RMSEP in §4.3.1.2) is ca. 0,7-1,2%. Doordat de set te klein is, kan ook voor de vleugels aan de hand van de set van het tweede seizoen geen acceptabel model opgesteld worden.

Tabel 4.10 Gegevens van de Dimorphotheca-vleugel-ijklijnen van het seizoen 1991 die het meeste perspectief bieden.

ijklijn	methode	RMSEC <sup>'91</sup>	RMSEP <sup>'90</sup>	bias	slope	RMSEP <sub>bc</sub>	RMSEP <sub>bsc</sub>
D23-4T	MLR	0,68	2,3	0,45	0,49	2,2	1,9
D2322-4T	MLR	0,35	1,5	0,65	0,88	1,3	1,3
D23-8	PLS	0,36	1,4	-0,41	0,78	1,3	1,2
D2322-4	PLS	0,43	1,3	-0,04	0,77	1,3	1,1

Na samenvoegen van de data (nieuwe set:  $n=80$ ) uit de beide seizoenen is opnieuw gerekend met zowel de hele ( $n=80$ ), als de gedeelde set ( $n=63 + n=17$ ). In figuur 4.7 (links) is het aantal PLS-factoren van de calibratieset ( $n=63$ ) uitgezet tegen de RMSEC, RMSEP en slope uit de voorspelling van de toetsset ( $n=17$ ). Omdat er al overfit optreedt bij het model met zeven factoren, is gekozen voor het model met zes (6) factoren. Tevens is in figuur 4.7 (rechts) de voorspelling opgenomen van de hele set ( $n=80$ ) met deze ijklijn DN6306.



Figuur 4.7    Het aantal PLS-factoren uitgezet tegen de RMSEC, RMSEP en slope uit de voorspelling van de toetsset (n=17,links). Oliegehalten van Dimorphotheca-vleugels bepaald met NIRS m.b.v. de ijklijn DN6306, voorspelling aan de hele set (n=58,rechts).

Het oliegehalte van Dimorphotheca-vleugels is te voorspellen met een RMSEP van ca. 0,6-0,8%, wat wil zeggen dat de nauwkeurigheid (95% betrouwbaarheidsinterval) van de voorspelling 1,2-1,6% is. In tabel 4.11 zijn de ijklijnen opgenomen die het meeste perspectief bieden. Voor de ijklijnen (n=80) moet worden vermeld dat de toetswaarden slechts indicatief zijn, omdat de toetsmonsters ook opgenomen zijn in de calibratieset. De bereikte nauwkeurigheid is iets beter dan uit de gecorrigeerde RMSEP (§4.3.1.2) verwacht mocht worden.

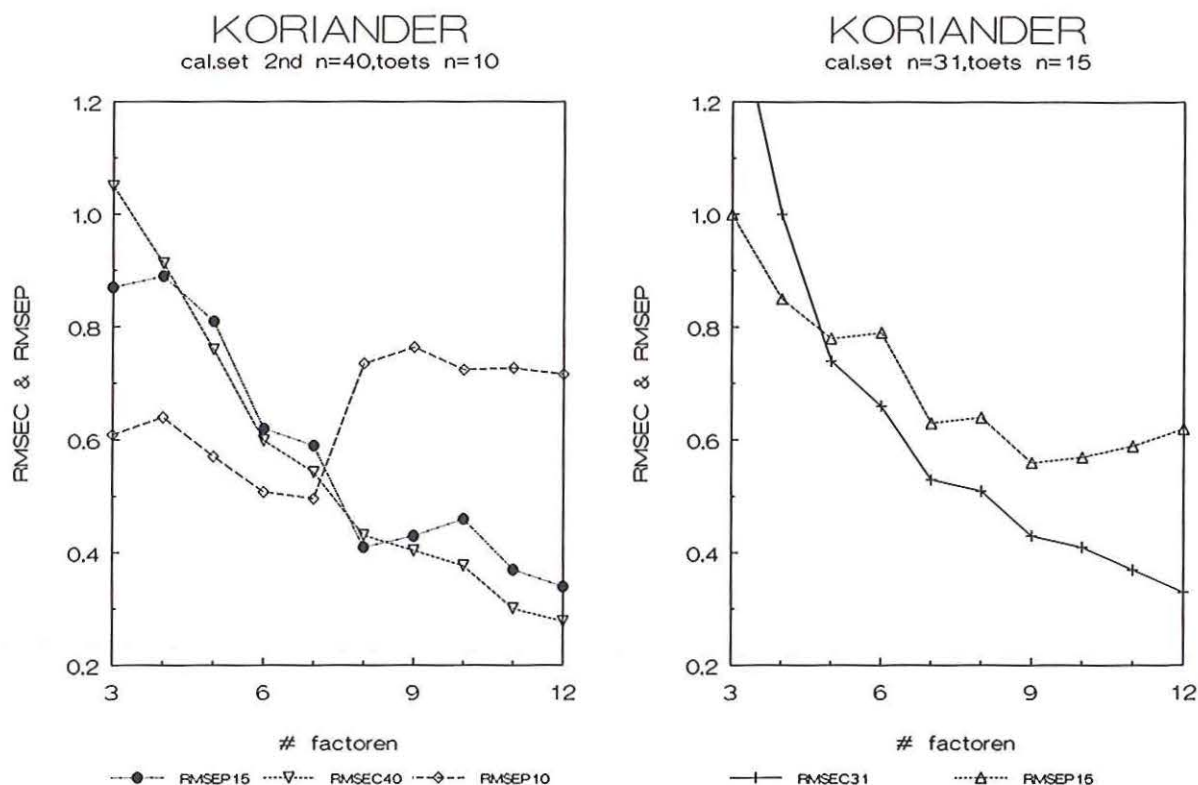


Tabel 4.11 Gegevens van de Dimorphotheca-vleugel-ijklijnen die het meeste perspectief bieden. Doordat de toetsset in de calibratie is opgenomen bij n-cal=80, zijn de waarden indicatief.

ijklijn	methode	RMSEC	R	n-cal	RMSEP <sup>17</sup>	Bias	Slope
D2ND63T4	MLR	0,95	0,95	63	0,80	0,07	0,972
D631-4TA	MLR	1,12	0,92	63	0,84	-0,08	0,912
D80208T4	MLR	0,91	0,95	80	0,76	0,03	0,979
DN6306	PLS	1,10	0,93	63	0,66	0,12	1,030
D2ND6307	PLS	0,75	0,97	63	0,64	0,08	0,955
D80-9	PLS	0,81	0,96	80	0,63	0,15	0,948
D802ND-6	PLS	0,76	0,97	80	0,57	0,08	0,984

#### 4.4 Coriandrum sativum

Van Coriander zijn slechts uit één seizoen (1991) monsters beschikbaar. De set bestaat uit 52 monsters, waarvan 2 monsters alleen gespleten korrels bevatten. Deze gespleten monsters zijn niet in de berekening meegenomen. De resterende dataset is voor de berekening gesplitst in een calibratieset (n=40) en een validatieset (n=10). In figuur 4.8 (links) is te zien dat er overfit optreedt na 7-PLS factoren. Bij nader onderzoek blijkt dat het model voor een belangrijk deel wordt bepaald door een beperkt aantal (4) monsters. Om een robuuster model te krijgen zijn deze monsters uit de berekening gelaten en is een meer evenwichtige verdeling gemaakt tussen calibratieset (n=31) en validatieset (n=15) (figuur 4.8, rechts). Gelet op het kleine aantal monsters en het relatief grote aantal factoren is gekozen voor 7 factoren (K3107) om overfit te voorkomen. In tabel 4.12 zijn de ijklijnen opgenomen die het meeste perspectief bieden. De monsters kunnen worden voorspeld met een RMSEP van ca. 0,6%. Dit wil zeggen dat de verwachte nauwkeurigheid van de voorspelling 1,2% (95% betrouwbaarheidsinterval) bedraagt.

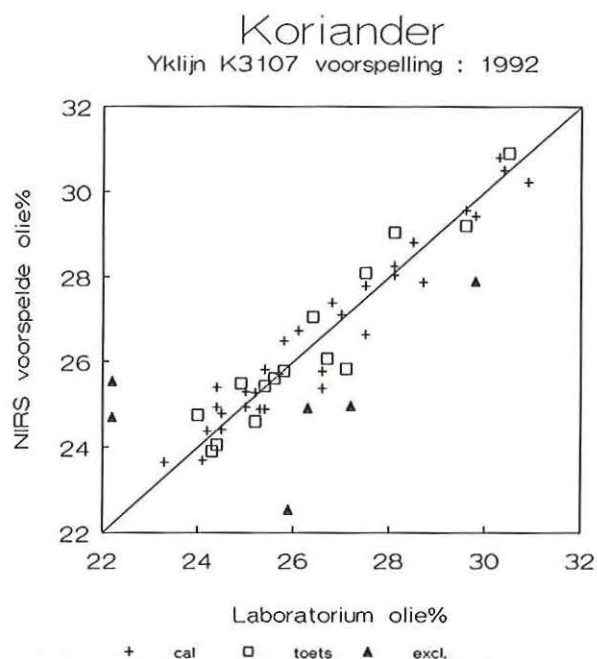


Figuur 4.8 Het aantal PLS-factoren uitgezet tegen de RMSEP ( $n=10$ , links  $n=15$ , rechts) en de RMSEC ( $n=40$ , links  $n=31$ , rechts) van koriander. De monsters die de modellen (links) onevenredig bepalen, zijn niet opgenomen in de nieuwe modellen (rechts).

Tabel 4.12 Gegevens van de Koriander-ijklijnen van het seizoen 1991.

ijklijn	methode	RMSEC	RMSEP <sup>15</sup>	RMSEP <sup>10</sup>	bias	slope
K31-2T	MLR	0,69	0,66	-	-0,16	0,914
K3107	PLS	0,53	0,63	-	0,02	0,888
K40-4T1F	MLR	0,79	-	0,50	0,04	1,014
K4007	PLS	0,69	-	0,72	-0,15	0,892
K40D07	PLS	0,54	-	0,50	0,02	1,028

In figuur 4.9 is de voorspelling met behulp van de ijklijn K3107 weergegeven van de hele korianderset ( $n=52$ ). Hierbij zijn de oliegehalten die klassiek zijn bepaald uitgezet tegen de met NIRS voorspelde oliegehalten. De monsters die de modellen onevenredig bepalen, zijn apart vermeld (excl).



Figuur 4.9 Voorspelling van de koriander-monsters van 1992 (totale set  $n=52$ ) aan de hand van de ijklijn K3107 ( $n=31$ ). De monsters die noch in de calibratieset (cal) noch in de toetsset (toets) zijn opgenomen, zijn apart vermeld (excl).

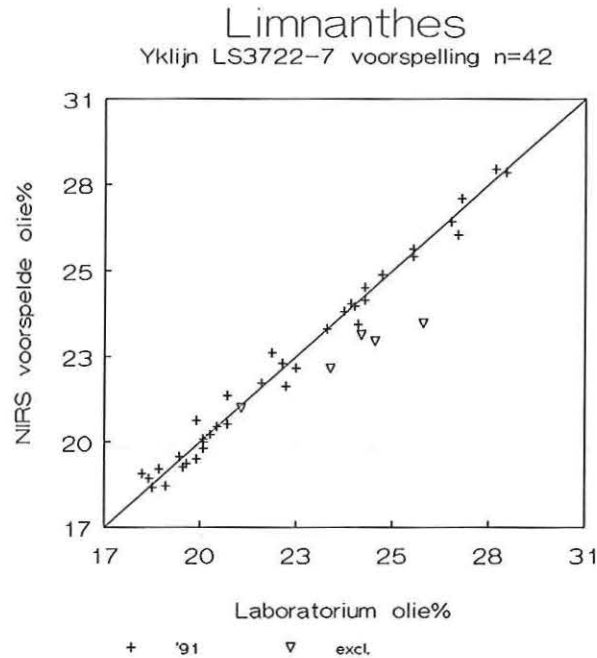
#### 4.5 Limnanthes alba

De dataset van *Limnanthes* is klein en onevenredig verdeeld over de seizoenen. De dataset bestaat uit 3 subgroepjes ; de PAGV-set (1991,  $n=18$ ) en de CPRO-sets (1991,  $n=19$  en 1986-1989,  $n=5$ ). De monsters ( $n=5$ ) uit de seizoenen '86 tot '89 blijken bij nader onderzoek het model onevenredig te beïnvloeden en zijn niet in de berekening meegenomen. De ijklijnen zullen in de praktijk moeten worden getoetst voor er een waarde-oordeel aan kan worden gegeven. In tabel 4.13 zijn de ijklijnen opgenomen die het meeste perspectief bieden. De verwachte nauwkeurigheid (95% betrouwbaarheidsinterval) van de bepaling is ca. 0,8% respectievelijk ca. 1,3% afhankelijk van de te gebruiken regressiemethode.

Tabel 4.13 Gegevens van de *Limnanthes*-ijklijnen van het seizoen 1991.

ijklijn	methode	RMSEC	R
L3722T4	MLR	0,62	0,98
LS3722-7	PLS	0,39	0,99

In figuur 4.10 is de voorspelling met behulp van de ijklijn LS3722-7 weergegeven van de hele Limnantheset (n=42). Hierbij zijn de oliegehalten die klassiek zijn bepaald uitgezet tegen de met NIRS voorspelde oliegehalten. De monsters die de modellen onevenredig bepalen, zijn apart vermeld (excl).



Figuur 4.10 Voorspelling van de Limnanthes-monsters van 1992 (totale set n=42) aan de hand van de ijklijn LS3722-7 (n=37). De monsters die niet in de calibratieset ('91) zijn opgenomen zijn apart vermeld (excl).

## 5 CONCLUSIES

- Het is gebleken dat de binnenmonstervariatie van oliehoudende zaden groot is. Gebruik van een grote (coarse) cuvet kan deze variatie uitmiddelen. Deelmonsters nemen met de Retsch-monsterverdeelmolen minimaliseert de variatie verder.
- De ijklijnen gebaseerd op één enkel seizoen (en/of grondsoort) zijn niet betrouwbaar voor de voorspelling van monsters uit een nieuw seizoen, doordat er verschillen in structuur en samenstelling zijn, die o.a. veroorzaakt worden door seizoensinvloeden, teeltverschillen, grondsoort, afrijpingsstadia.
- De verwachte nauwkeurigheid (95% betrouwbaarheidsinterval) van de voorspelling van het oliegehalte van monsters uit een volgend seizoen is voor *Crambe abyssinica*, *Dimorphotheca pluvalis* en *Coriander sativum* ca. 1,2%, voor *Euphorbia lagascae* ca. 0,9% en voor *Limnanthes alba* ca. 0,8%. Deze nauwkeurigheden zijn voor screeningsdoeleinden voldoende nauwkeurig.



- Berekeningen met PLS kunnen bij meer complexe matrices een beter resultaat opleveren dan bij MLR, maar gemiddeld genomen is het resultaat gelijk. Vanwege de collineariteitsproblemen die bij MLR kunnen optreden en vanwege de snelheid van calibreren gaat de voorkeur uit naar met PLS geconstrueerde ijklijnen.
- Aanbevolen wordt de voorspelmodellen (ijklijnen) jaarlijks op nauwkeurigheid te controleren en uit te breiden met een representatieve set monsters uit het nieuwe oogstseizoen.

N.B. De gekozen ijklijnen van de verschillende zaadtypen (waar mogelijk getoetst), waarvan verwacht wordt dat deze het meest robuust zijn, zijn opgenomen in tabel B.1 in bijlage B.

## LITERATUUR

Draper, N.R. & Smith, H.

Applied regression Analysis

Wiley, New York, USA (1981)

Genstat

Genstat 5 reference manual version 2.2

Clarendon Press, Oxford, UK (1987)

Kamp, H.J. van der; Groenendijk, F.P.F.; Oostrom, J.J. van; Frankhuizen, R.

Oriënterend onderzoek naar de bepaling van het oliegehalte in *Dimorphotheca pluvialis*, *Crambe abyssinica* en *Euphorbia lagascae* met NMR en NIRS.

Rapport 91.52, RIKILT-DLO, Wageningen (1991)

Kamp, H.J. van der; Oostrom, J.J. van; Groenendijk, F.P.F.

Bepaling van het oliegehalte van *Crambe abyssinica*, *Euphorbia lagascae* *Dimorphotheca pluvialis* en *Limnanthes alba* met lage resolutie NMR.

Rapport 92.19, RIKILT-DLO, Wageningen (1992)

Martens, H. & Naes, T.

Multivariate calibration

Wiley, Chichester, UK (1989)

Montgomery, D.C. & Peck, E.A.

Introduction to linear regression analysis

Wiley, New York, USA (1982)

NSAS

Manual for Near infrared Spectral Analysis Software (NSAS), version 3.25

NIRSystems Inc., SilverSpring, USA (1989)

Unscrambler

Unscrambler II User's guide, extended memory version 4.10

CAMO A/S, Trondheim, Norway (1993)



## Bijlage A

### GEBRUIKTE STATISTISCHE FORMULES

RMSEC (Root Mean Square Error of Calibration) = standaarddeviatie van de verschillen tussen de met NIRS voorspelde en met de klassieke methode gevonden gehalten van de calibratieset.

$$RMSEC = \frac{\sqrt{\sum (x-y)^2}}{(n-1)}$$

RMSEP (Root Mean Square Error of Prediction) = standaarddeviatie van de verschillen tussen de met NIRS voorspelde en met de klassieke methode gevonden gehalten van de toetsset.

$$RMSEP = \frac{\sqrt{\sum (x-y)^2}}{(n-1)}$$

r = correlatiecoëfficiënt van de regressielijn  $Y=aX+b$

$$r = \frac{\sum (x_i y_i - n \bar{x} \bar{y})}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}}$$

Slope = hellingshoek van de regressielijn t.o.v. de 45° lijn.

$$Slope = \frac{n \sum (yx) - \sum y \sum x}{n \sum y^2 - (\sum y)^2}$$

Bias = gemiddeld verschil tussen de met nirs voorspelde gehalten en de klassiek gevonden gehalten.

$$Bias = \frac{\sum (y-x)}{n}$$

waarin      x = klassiek gevonden gehalten  
              y = NIRS voorspelde gehalten  
              n = aantal monsters

## Bijlage B

Tabel B.1 De gekozen ijklijnen van de verschillende zaadtypen waarvan verwacht wordt dat deze het meest robuust zijn.

type oliehoudend zaad	MLR		PLS	
	ijklijn	RMSEP	ijklijn	RMSEP
<i>Crambe abyssinica</i>	C4N2T3F3	0,54	CN44G1P5	0,52
<i>Euphorbia lagascae</i>	E116T3F2	0,40	E11607	0,43
<i>Dimorphotheca pluvialis</i> kegels	DK90914T	0,67	DK44-8	0,69
<i>Dimorphotheca pluvialis</i> vleugels	D2ND63T4	0,80	D2ND6307	0,64
<i>Corandrum sativum</i>	K31-2T	0,66	K3107	0,63
<i>Limnanthes alba</i>	L3722T4	0,62	LS3722-7	0,39



